MRC | Epidemiology Unit

Design and Analysis for Genetic Study of Complex Traits

October 20-26, 2008

Shandong University, Jinan, China

Jing Hua Zhao

MRC Epidemiology Unit, Cambridge, UK

jinghua.zhao@mrc-epid.cam.ac.uk
http://www.mrc-epid.cam.ac.uk/~jinghua.zhao

Aim

- To give an overview of genetic study of complex traits
- To focus issues in study design and statistical analysis
- To show case studies
- To have hands-on exercise
- …
- To have fruitful discussions and your feedback!

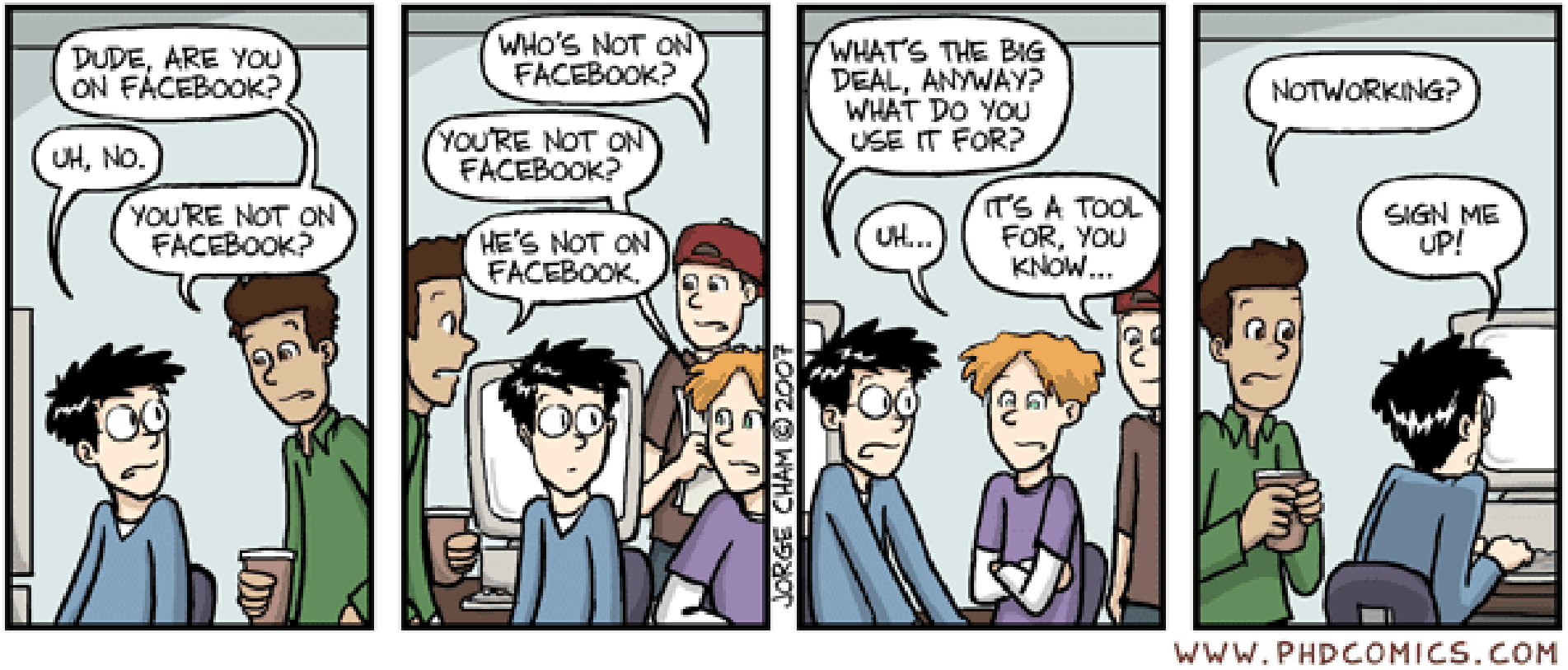
Course outline

- It was intended as a weekly lecture series
- Most sessions last for two hours of lectures and practice (including a break)

1. Introduction
2. Analysis of family resemblance and segregation
3. Linkage analysis
4. Association analysis
5. Issues in association analysis
6. Study design
7. Advanced topics and summary

Self-Introduction

Yourself and your expectation



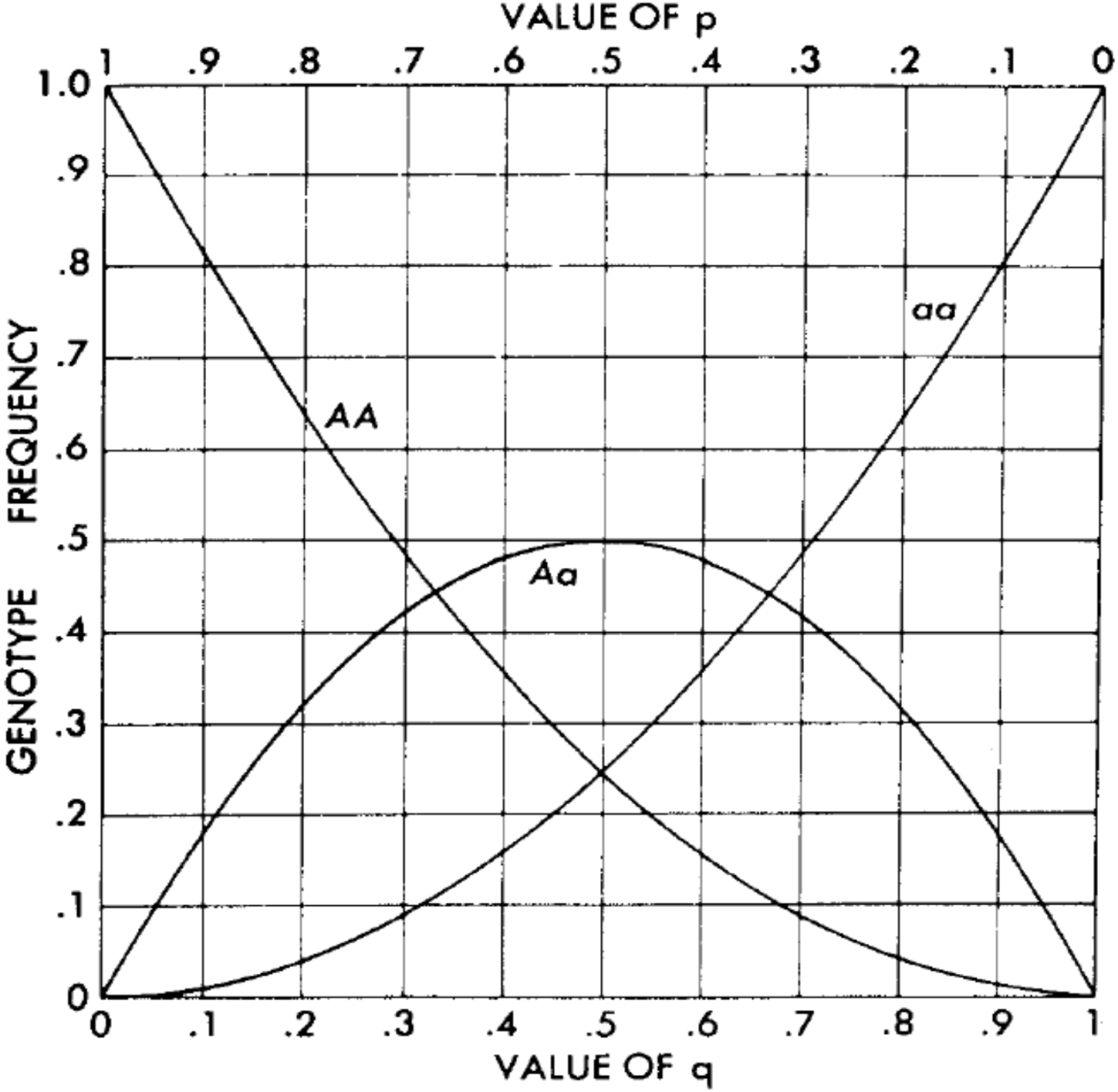
We are more positive than this!

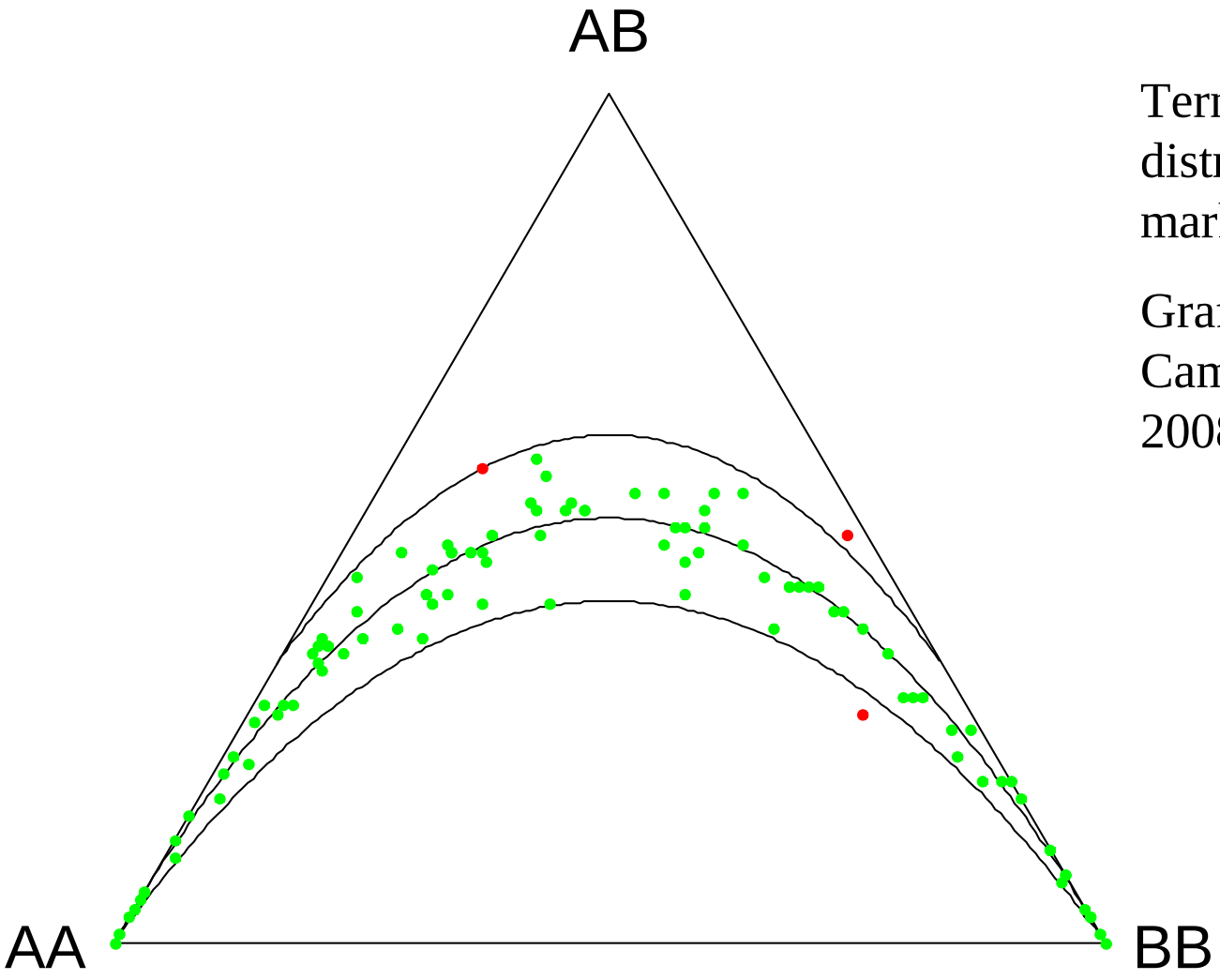
Genetics, epidemiology (King et al. 2006)

- **Genetics** is the scientific discipline dealing with 1. the study of inheritance and variation of biological traits, and 2. the study of genes, including their structure, function, variation, and transmission. **Heredity** is the passing of traits to offspring.
- **Population genetics** is the study of the genetic composition of populations. **Population geneticists** try to estimate gene frequencies and detect the selective influences that determine them in natural populations. They also build mathematic models to elucidate the interaction of factors such as selection, population size, mutation, and migration upon the fixation and loss of linked and unlinked genes.
- **Epidemiology** is the study of factors affecting the health and illness of populations, and serves as the foundation and logic of interventions made in the interest of public health and preventive medicine (http://en.wikipedia.org/wiki/Epidemiology).

Concepts in population genetics

- **Random genetic drift**: For each generation there is an element of chance in the drawing of gametes that will unite with other to form the next generation. Chance alone can result in changes in allele frequency which does not follow any predetermined direction.
- **Mutation**: a change in DNA sequence which provides new variation.
- **Migration**: movement between populations decreases divergence.
- **Selection**: as differential survival or fertility of genes of certain allelic types may increase frequency of the haplotype containing the allele.
- **Assortative mating**: mate choice is influenced by genotype of phenotype, e.g., positive when mates are more like each other.
- **Hardy-Weinberg equilibrium**: the concept that both gene frequencies and genotype frequencies will remain constant form generation to generation in an infinitely large, interbreeding population in which mating is at random and there is no selection, migration, or mutation (see relationships between frequencies of genes A or a and the genotype frequencies AA, Aa, aa predicted by HWE).


VALUE OF p
1
.9
.8
.7
.6
.5
.4
.3
.2
.1
0
1.0
.9
.8
.7
.6
.5
.4
.3
.2
.1
0
GENOTYPE FREQUENCY
AA
Aa
aa
0
.1
.2
.3
.4
.5
.6
.7
.8
.9
1
VALUE OF q



Ternary plot showing distributions of 100 markers for 100 SNPs

Graffelman & Morales-Camarena *Hum Hered* 2008

Genetic epidemiology (Wikipedia)

The study of the role of genetic factors in determining health and disease in families and in populations, and the interplay of such genetic factors with environmental factors. Slightly more formally, genetic epidemiology was defined by Morton as "*a science which deals with the etiology, distribution, and control of disease in groups of relatives and with inherited causes of disease in populations*" (Morton 1982).

It is closely allied to both molecular epidemiology and statistical genetics, but these overlapping fields each have distinct emphases, societies and journals. The former is a branch of public health that deals with the contribution of potential genetic and environmental risk factors identified at the molecular level, to the etiology, distribution and control of the disease in groups of relatives and populations. It therefore improves our understanding of the pathogenesis of disease by identifying specific pathways, molecules and genes that influence the risk of developing disease.

Scopes of genetic epidemiology

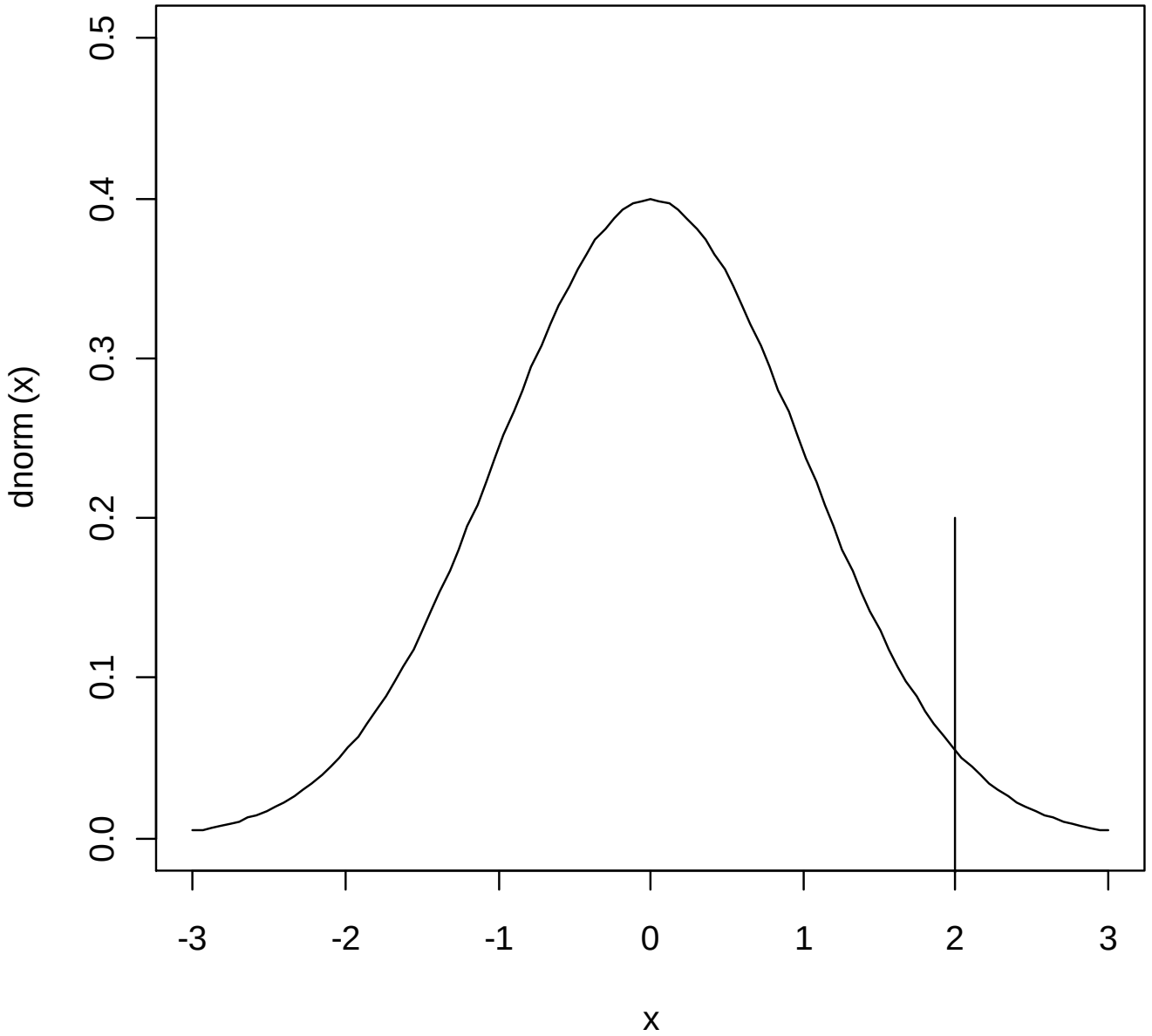
- It traditionally studies the role of genetics in disease progression using several study designs: 1. Familial aggregation: Is there a genetic component to the disease, and what are the relative contributions of genes and environment? 2. Segregation: What is the pattern of inheritance of the disease (e.g. dominant or recessive)? 3. Linkage: On which part of which chromosome is the disease gene located? 4. Association: Which allele of which gene is associated with the disease?
- It has proved highly successful for locating genes for monogenic disorders and the recent focus is on complex traits for which many genes are involved (polygenic, multifactorial or multigenic disorders). This has developed rapidly following completion of the Human Genome Project, as advances in genotyping technology enables genome-wide association studies (GWAS) of single nucleotide polymorphisms in thousands of individuals. These have lead to the discovery of many genetic polymorphisms that influence the risk of developing many common diseases.

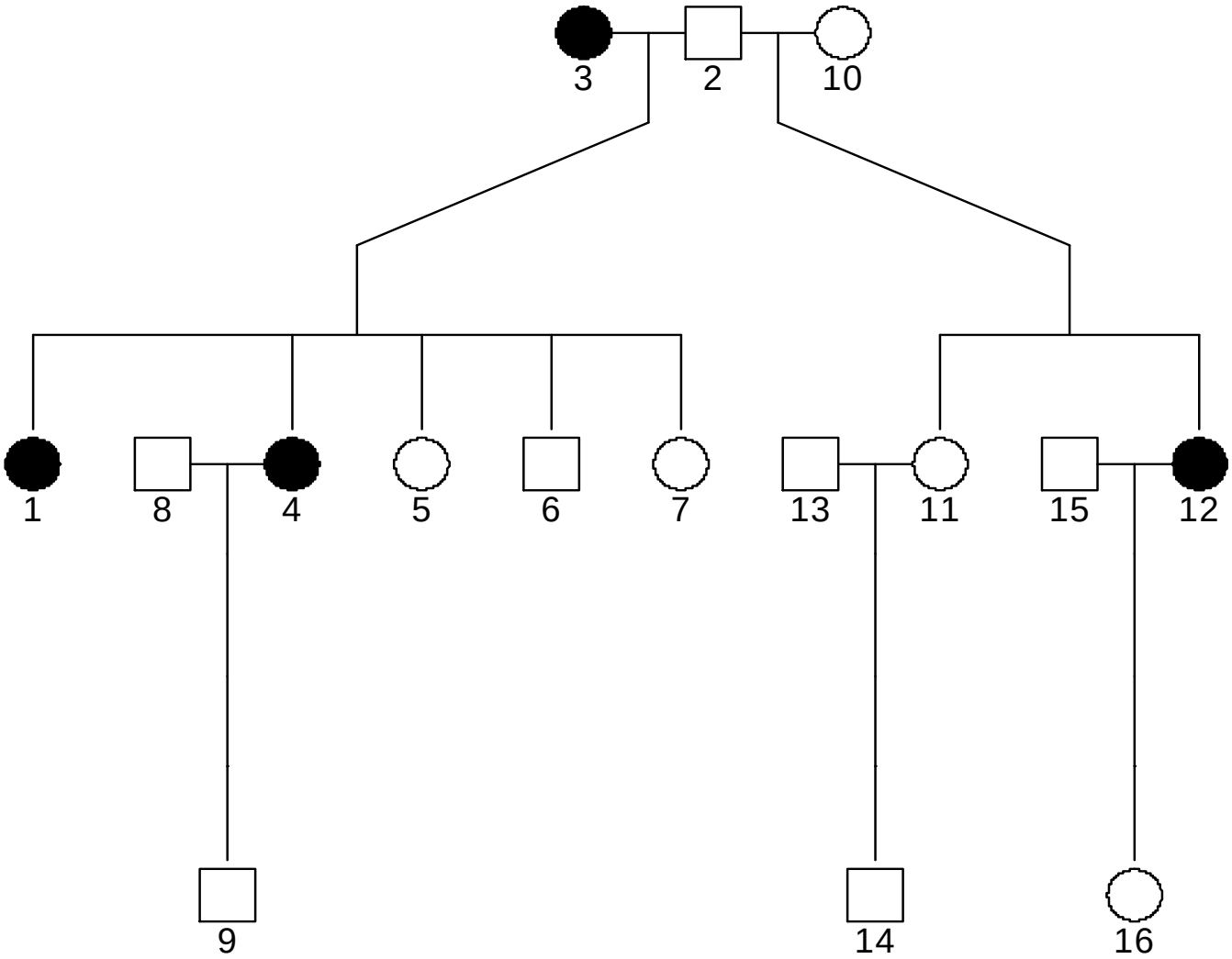
We need some background

- Basic terminology:
  - Population and family
  - Chromosome, DNA, genotype, phenotype, genes, genetic markers
- Some technology: Genotyping, PCR, GeneChips
- Collaborative projects: Human Genome Project, HapMap, 1000 genomes

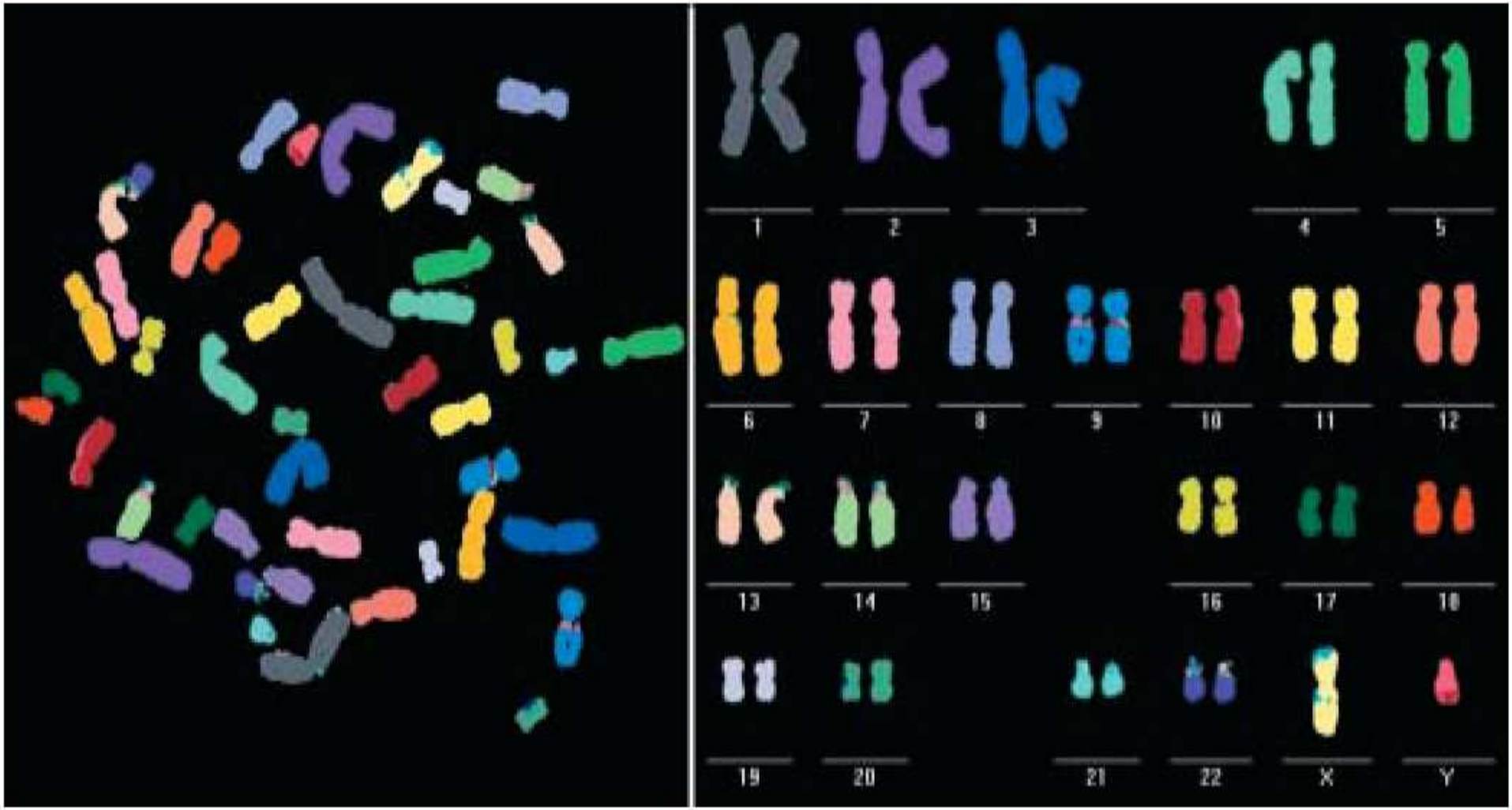
Genotype, phenotype and their relationship

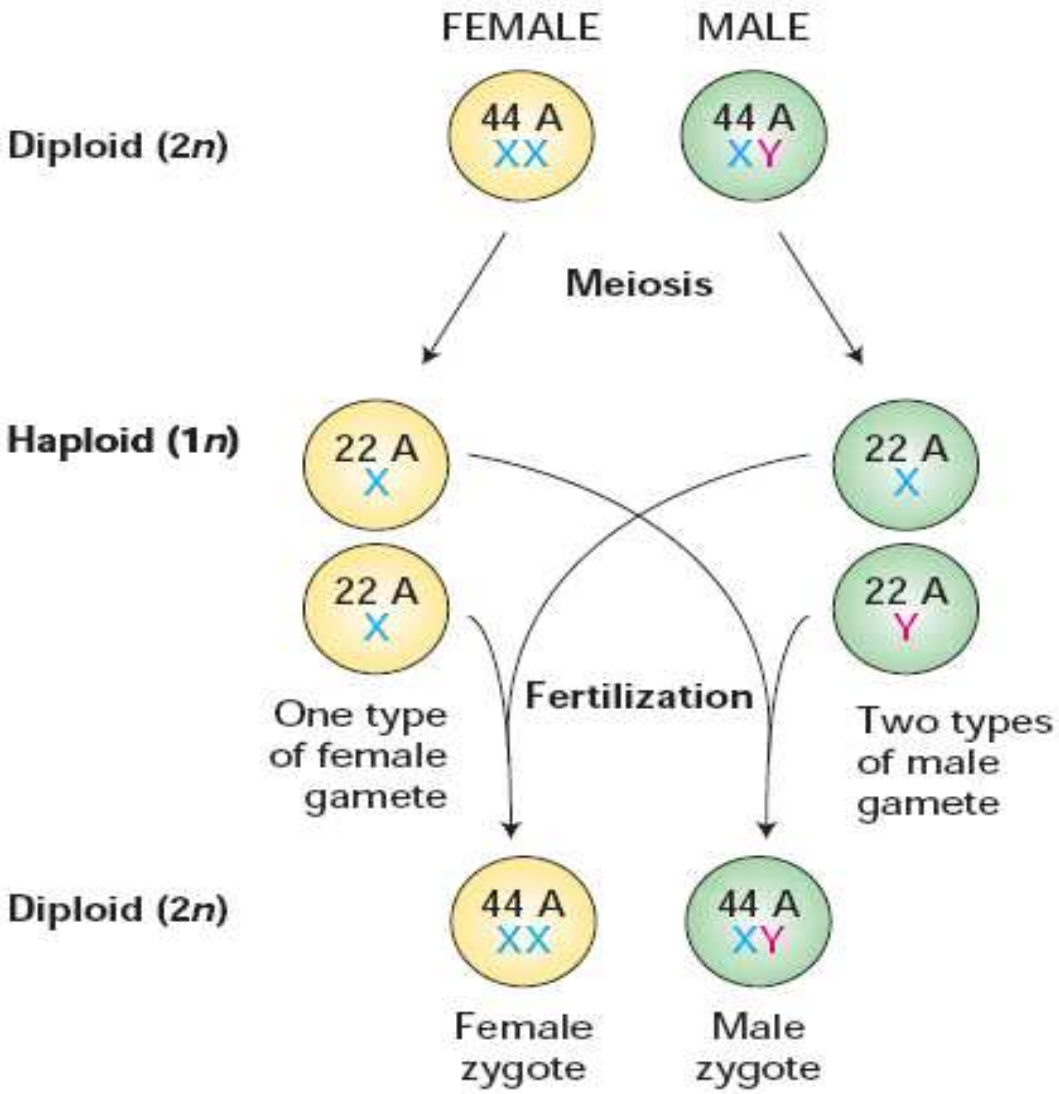
- Genome
- Chromosome: autosomes (22) + XX/XY (1)
- DNA: 3E+9 nucleotides, mtDNA ~17000
- Gene, locus, allele, polymorphism, marker, homozygous and heterozygous
- CNV
- Genotype (G), phenotype (P), haplotype (H)
- Penetrance: f(P|G)
- Pedigrees, founders/non-founders
- Mendelian laws of inheritance
  - A parent transmits one allele of a gene chosen at random to offspring
  - Alleles from any two genes transmitted independently to offspring. However, this is true only for non-syntenic loci


dnorm (x)
0.0
0.1
0.2
0.3
0.4
0.5
-3
-2
-1
0
1
2
3
x


3
2
10
1
8
4
5
6
7
13
11
15
12
9
14
16

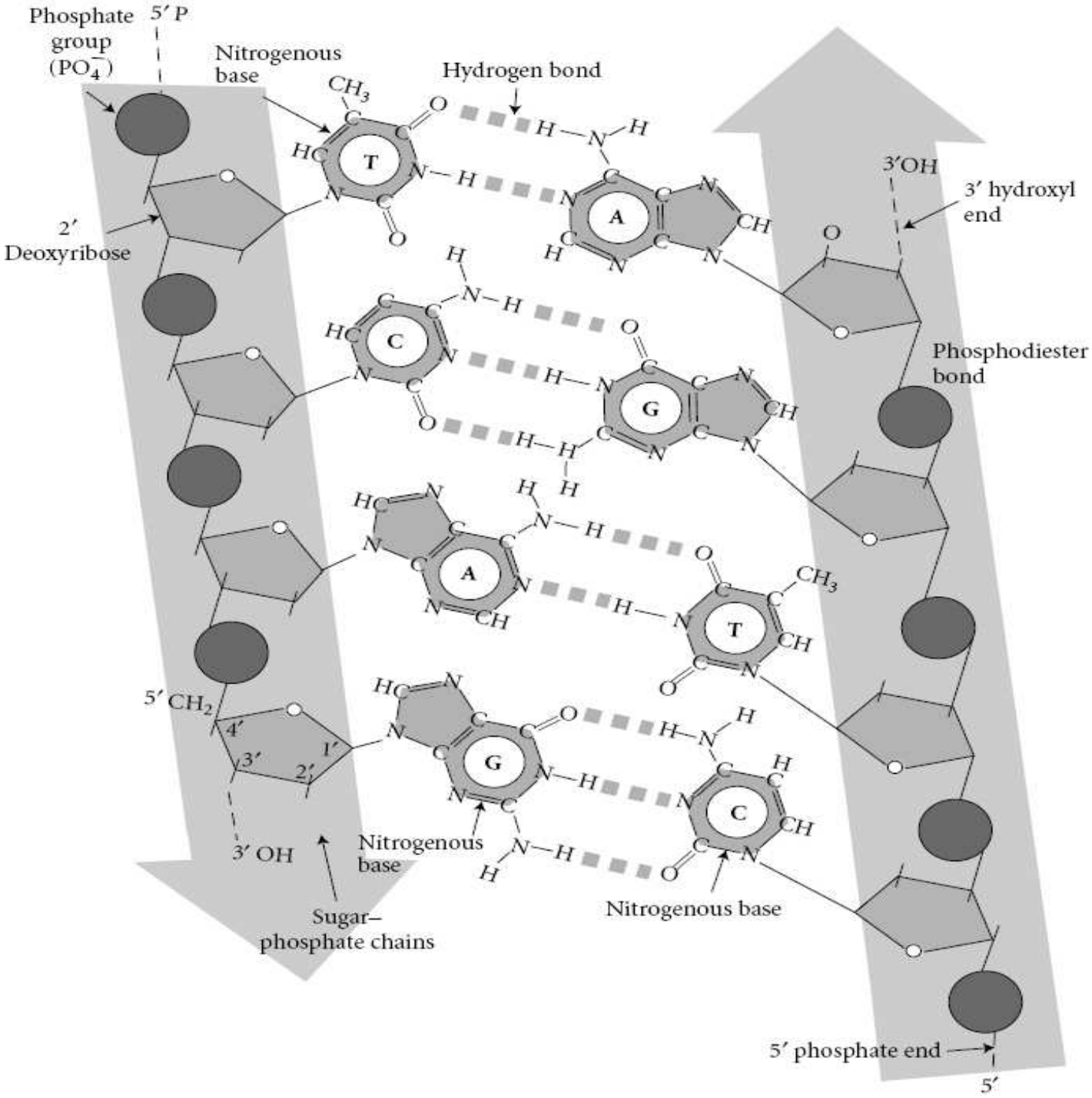
23 pairs of distinct chromosomes




FEMALE
MALE
Diploid (2n)
44 A
XX
44 A
XY
Meiosis
Haploid (1n)
22 A
X
22 A
X
22 A
X
22 A
Y
Fertilization
One type
of female
gamete
Two types
of male
gamete
Diploid (2n)
44 A
XX
44 A
XY
Female
zygote
Male
zygote

Non-independence of segregation

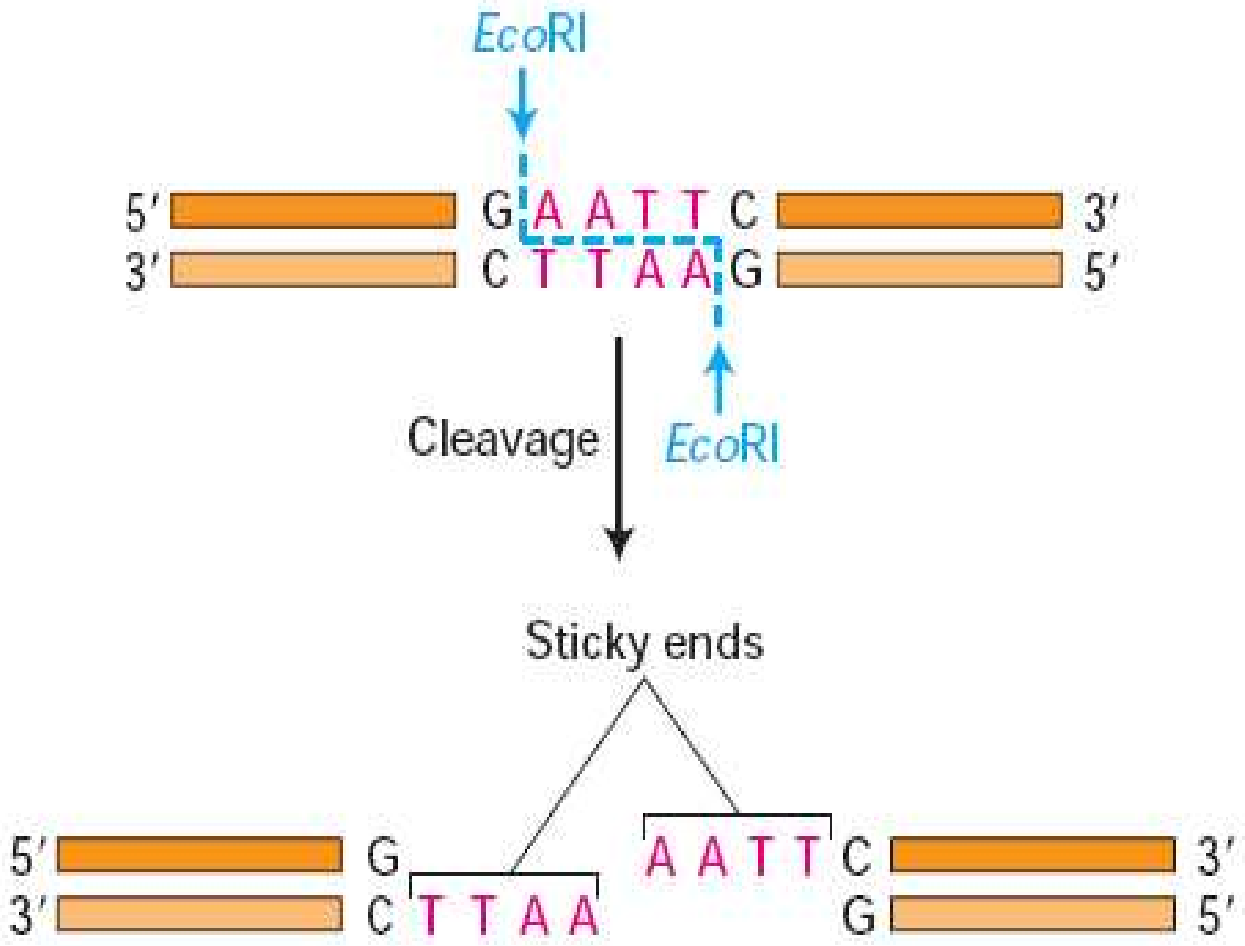
- **Linkage**: genes on the same chromosome tend to be inherited together
- **Meiosis**: each member of pair of homologous chromosomes replicates to form two sister chromosomes known as chromatiids
- **Cross-over**: reciprocal exchange of segments between non-sister chromatids during meiosis
- **Interference**: lack of independence in recombinations at different intervals on a chromosome


Phosphate group (PO_4^-)
5′ P
Nitrogenous base
Hydrogen bond
2′ Deoxyribose
3′OH
3′ hydroxyl end
Phosphodiester bond
5′ CH_2
4′
1′
3′
2′
3′ OH
Nitrogenous base
Nitrogenous base
Sugar–phosphate chains
5′ phosphate end
5′
T
A
C
G
A
T
G
C

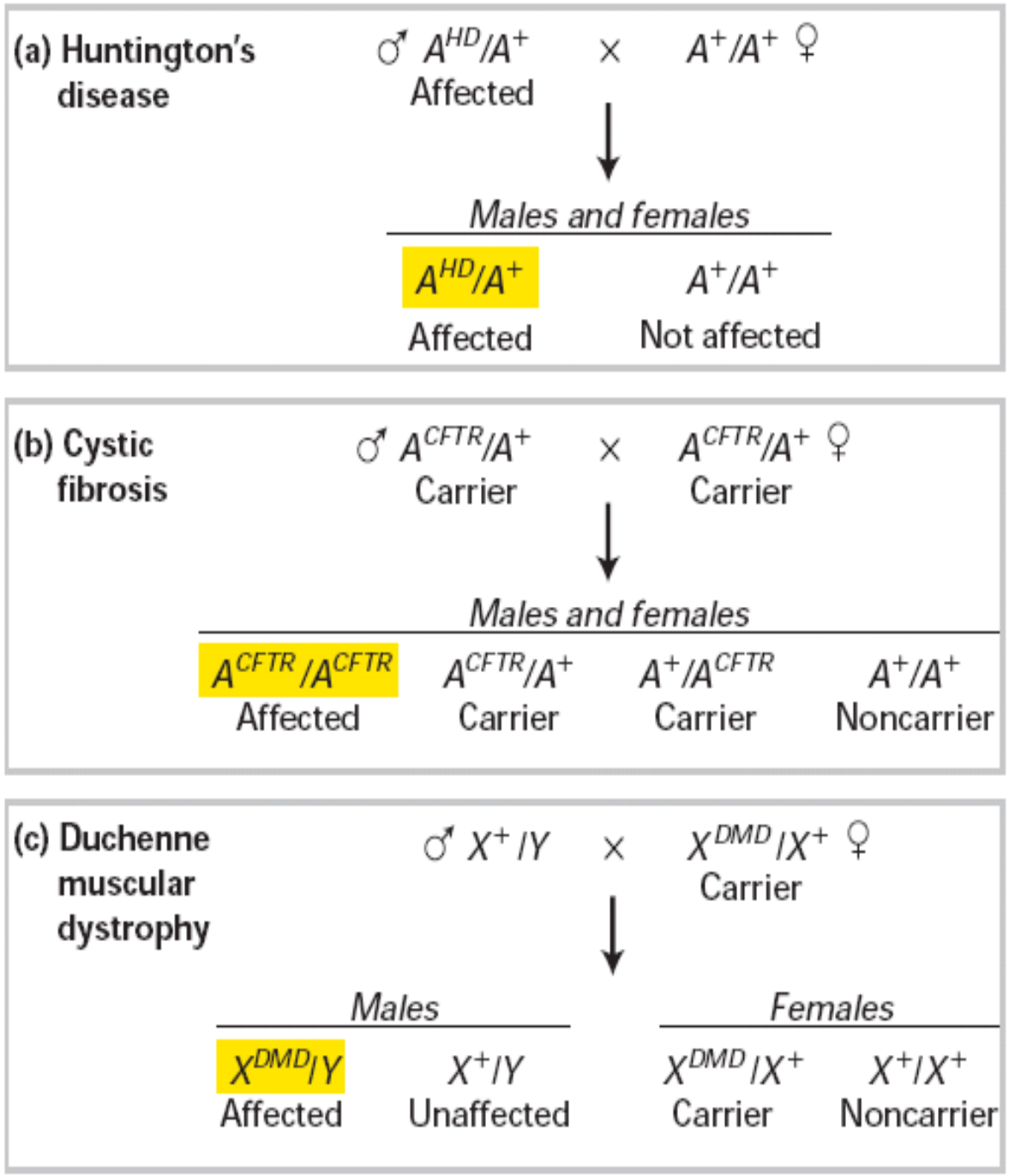
Nucleic acids (Serre 2006)

- DNA and RNA molecules are linear molecules made up of a succession of nucleotides. A nucleotide consists of a sugar (pentose with five carbons numbered 1 to 5) carrying a phosphate on its 5 carbon and a nitrogenous base on its 1 carbon. RNA contains a ribose while DNA contains a deoxyribose (no hydroxyl on the 2 carbon.
- Each deoxyribose of the DNA carries one of the four following bases: adenine (A), guanine (G), cytosine (C) and thyamine (T); for RNA, each ribose carries one of the following bases: adenine (A), guanine (G), cytosine (C) and uracil (U, acting in place of the thyamine). Another nitrogenous base is sometimes present but at a very low frequency.
- Nucleotides are linked by a phosphodiester (sugar-phosphate) bond between the 5 phosphate of one nucleotide and the 3 hydroxyl of the previous nucleotide. DNA and RNA are synthesized from the 5 to the 3 end in order to keep the 5P and the 3OH extremities in the same orientation.

Restriction enzymes are endonucleases produced by bacteria that typically recognize specific 4- to 8-bp sequences, called *restriction sites,* and then cleave both DNA strands at this site. Restriction sites commonly are short *palindromic* sequences; that is, the restriction-site sequence is the same on each DNA strand when read in the 5'→ 3' direction.



This restriction enzyme from *E. coli* makes staggered cuts at the specific 6-bp inverted repeat (palindromic) sequence shown, yielding fragments with single-stranded, complementary "sticky" ends. Many other restriction enzymes also produce fragments with sticky ends.



Autosomal dominant

Autosomal recessive

X-linked recessive

Complex traits

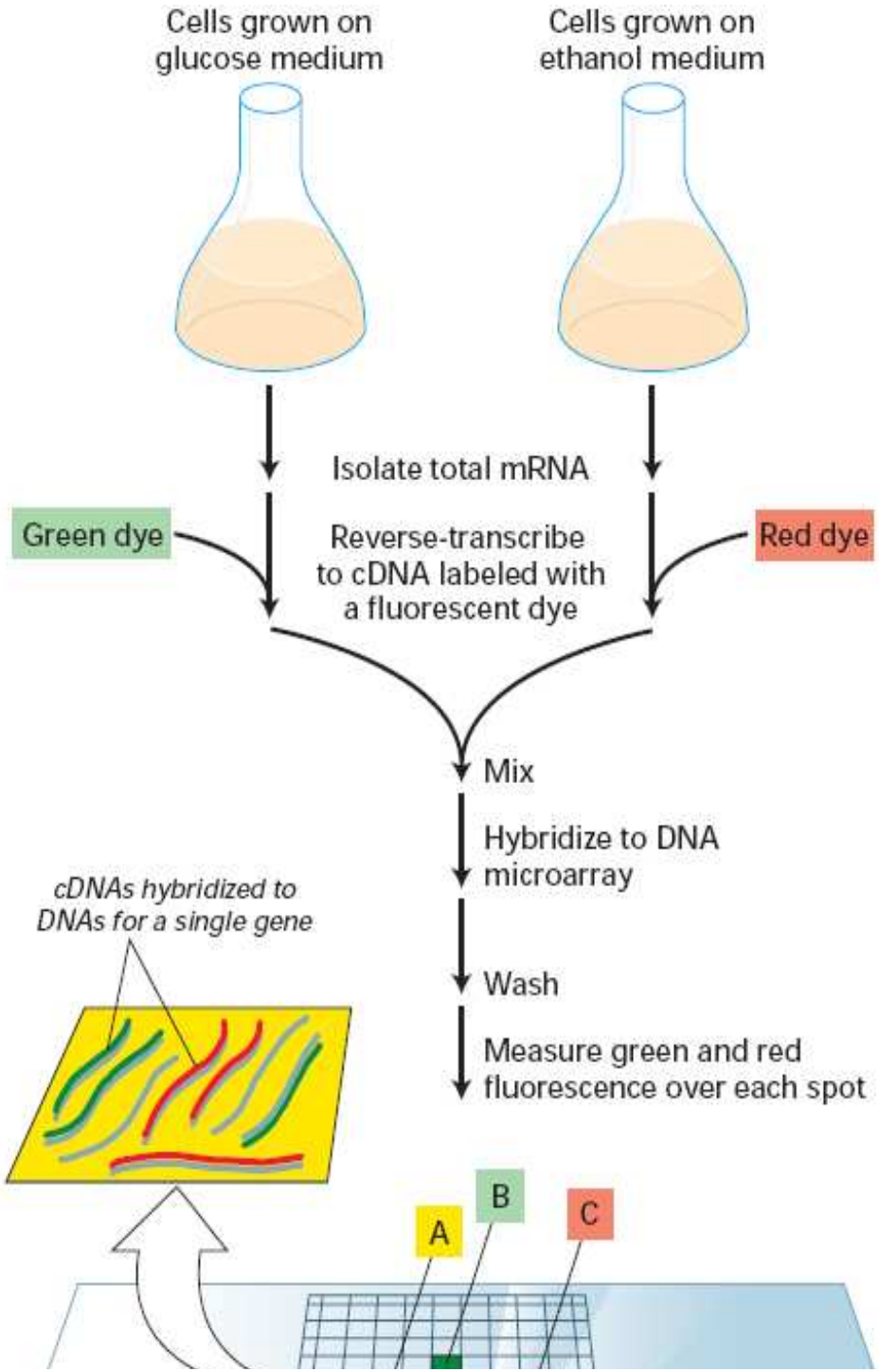
- **Complex traits**: the complex aetiology is characterized by
- **Incomplete penetrance**: the probability that individuals inheriting the gene will have the disease is less than 1 and dependent on factors such as age
- **Phenocopy**: a disease due to non-genetic causes
- **Genetic heterogeneity**: a disease due to different genetic mutations to different individuals
- **Polygenic inheritance**: the liability of disease due to additive or interactive effects at multiple loci
- Other phenomena such as **imprinting**, the difference in function of an allele according to its parental origin

If a spot is yellow, expression of that gene is the same in cells grown either on glucose or ethanol

If a spot is green, expression of that gene is greater in cells grown in glucose

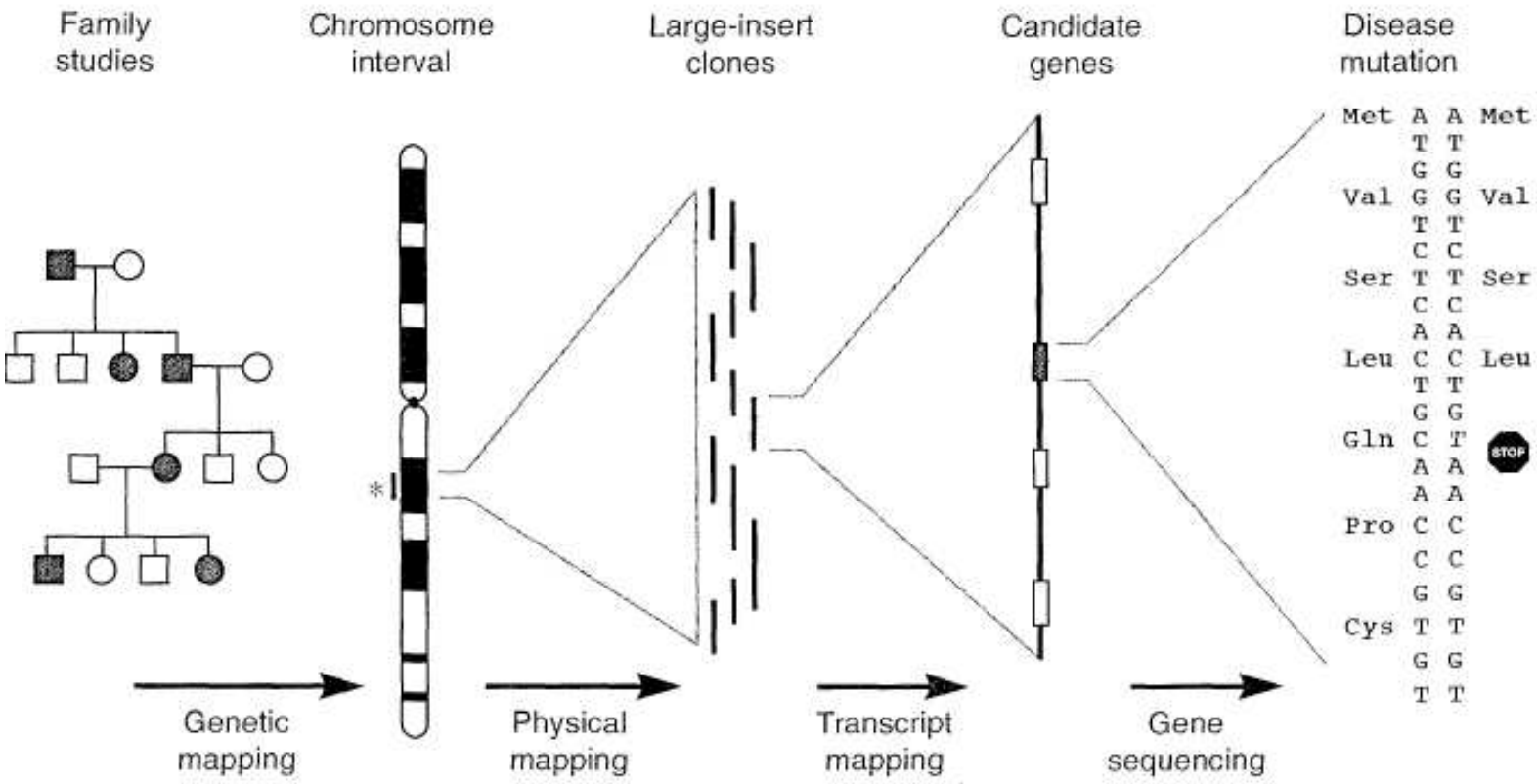
If a spot is red, expression of that gene is greater in cells grown in ethanol

(Lodish et al. (2003))



Some statistical terminology

- **Additive**: accumulative effects of individual alleles
- **Dominance**: interaction between alleles at the same locus
- **Epistasis**: interaction between alleles at different loci
- **Pleiotropic expression**: a situation in where an allele has more than one distinct phenotypic effect


Family studies
Chromosome interval
Large-insert clones
Candidate genes
Disease mutation
Met A A Met
T T
G G
Val G G Val
T T
C C
Ser T T Ser
C C
A A
Leu C C Leu
T T
G G
Gln C T
A A
A A
Pro C C
C C
G G
Cys T T
G G
T T
STOP
*
Genetic mapping
Physical mapping
Transcript mapping
Gene sequencing

Principles in dissecting gene component

- **Positional cloning**: to map and identify previously unknown loci
  - To develop a disease model for model of inheritance
  - To localize genes to a chromosomal region across the whole genome
- **Candidate gene**: To study loci which are suspected to play a key role in the disease
- There are two major difficulties in genetic study of human diseases: (1) one cannot arrange matings at will but rather must retrospectively interpret existing families; (2) the trait may not be simply related to the genotype at a single gene.
- Even with dense map of DNA polymorphisms, one cannot distinguish between homologus chromosome for homozygous individuals. There may also be difficulty for individuals heterozygous at a gene when parental information is unknown.

History ~1900

- 1859: It saw the first printing of Charles Darwin’s book, On the origin of Species by Means of Natural Selection, or the preservation of Favoured Races in the Struggle for Life.
- 1865: Gregor Mendel described experiments with peas showing heredity is transmitted in discrete traits.
- 1869: Friedrich Miescher isolates DNA (nuclein which is rich in phosphorus) through study of white blood cells from pus on bandages, as well as salmon sperm
- 1879: Walter Flemming described chromosome behavior during animal cell division through salamander embryos and stain chromosome
- 1900: Three botanists - Hugo DeVries, Carl Correns and Erich von Tschermak - independently rediscovered Mendel's work in the same year and helped expand awareness of the Mendelian laws of inheritance in the scientific world.

~1955

- 1911:Thomas Hunt Morgan and his group at Columbia University showed that genes, strung on chromosomes, are the units of heredity, His students, who included Alfred Sturtevant, Hermann Muller and Calvin Bridges, studied the fruit fly *Drosophila melanogaster*. They showed that chromosomes carry genes, discovered genetic linkage - the fact that genes are arrayed on linear chromosomes - and described chromosome recombination.
- In 1933, Morgan received the Nobel Prize in Physiology or Medicine for helping establish the chromosome theory of inheritance.
- 1953: Francis Crick and James Watson described the double helix structure of DNA.
- 1955: Joe Hin Tjio defined 46 as the exact number human chromosomes

~2005 (… more from www.genome.org and king et al. 2006)

- 1966:Genetic code cracked.
- 1977: DNA sequencing by Sanger and his colleagues, and Maxam and Gilbert developed rapid DNA sequencing methods
- 1983: PCR invented.
- 1989: Microsatellites as markers in the genome and detectable through PCR
- 1990: Launch of the Human Genome Project: mapping the human genome and eventually determining the sequence of all 3.2 billion letters in it; mapping and sequencing the genomes of other organisms important to the study of biology; developing technology for analyzing DNA; and studying the ethical, legal and social implications of genome research.
- 1996: Human gene map created.
- 2001: first draft of the Human Genome Sequence.
- 2005: HapMap Project completed.

A statistical perspective

- ~1950: map function, allele-sharing method, quantitative traits by Francis Galton, Karl Pearson, and RA Fisher
- 1950s: Lod score method
- 1970s: Elston-Stewart algorithm, Haseman-Elston on QTL, LIPED
- 1980s: PATHMIX, POINTER, LINKAGE, Lander-Green algorithm, MENDEL, FISHER
- 1990s: TDT, GENEHUNTER, SAGE, SOLAR, Loki
- 2000s: haplotype analysis, GWAS on epidemiological cohorts, Allegro, Merlin, Superlink, Haploview, PLINK, IMPUTE/MACH
- Algol, Fortran (1950s), C, Pascal (1970s), C++ (1980s)
- SAS (1970s), Stata (1980s), R (1990s)
- Morton (1982), Morton et al. (1983), Khoury et al. (1993), Thomas (2004)

Therefore we do not cover

- **Evolution**, the process of change in the inherited trait of a population from one generation to the next.
- **Epigenetics**, the study of the mechanisms by which genes bring about their phenotypic effects (King et al. 2006), or the heritable changes in gene expression caused by mechanisms other than changes in the underlying DNA (http://en.wikipedia.org/wiki/Epigenetics).
- **Epigenotype** is the series of interrelated developmental pathways through which the adult form is realized.
- Sequence analysis.
- Proteomics.
- …

Mathematical Statistics

- Random variables
- Distribution
- Mean, variance, covariance, correlation
- Significant testing and type-I and type-II errors
- Descriptive analysis
- Statistical model
  - Distribution (Normal and Chi-squared)
  - Linear, general linear, generalised linear, mixed models
- Likelihood
  - Fisher's information
  - Wald's, score, and log-likelihood ratio tests
- Bayesian inference
- Meta-analysis
- Multiple imputation
- Multivariate analysis (PCA, MDS)

Statistical inference and statistical tests

- **Statistical inference**: parameter estimation and hypothesis testing; the optimality of a statistical test can be assessed by the following criteria:
- **Validity**: the correct type I error under the null hypothesis
- **Power**: the probability of rejecting the null hypothesis when an alternative is true
- **Robustness**: the resistance to departure from the idealized assumptions

Statistical computing

- Computer systems (Windows, Linux)
- High-level programming languages (Fortran, C/C++)
- Specialized software
  - PATHMIX, POINTER, SAGE
  - Linkage: LINKAGE, GENEHUNTER, Merlin, SOLAR
  - Association: EIGENSTRAT, IMPUTE/SNPTEST, PLINK
- General systems: SAS, Stata, R, M*plus*
  - ODBC
  - Internet access

Break

Evolution of computer implementations

- Algorithms evolution span over ~50 years, with a list of software programs (Pascal, Fortran, C/C++,…), (http://inkage.rockefeller.edu). Increasingly limited and requires marriage with established systems (commercial and free, e.g. SAS/Stata/S-PLUS and R).
- In particular for R: The call for integration is reminiscent of the development of Linux system. R is a general computing environment for both practitioners and programmers. Standard statistical and genetic analyses can be conducted in a reliable and unified fashion. It makes efficient statistical modelling possible. It is accessible and free. It is easy to extend and a large number of packages maintained by a large and dedicated team. Zhao & Tan. *Hum. Genomics*, **2**, 258-265, 2006

SAS

- A system (http://www.sas.com, http://en.wikipedia.org/wiki/SAS) with modelling facilities, statistics, operations research, with database support such as Oracle, MySQL and facility such as ODBC, data visualisation, fantastic data subsetting facility e.g.,

```
PROC SQL;
    CREATE TABLE mytable AS SELECT * FROM  master
    WHERE rsn IN (SELECT rsn FROM d1) & id IN (SELECT
    id from d2);
QUIT;
```

- SAS/Genetics which includes procedures ALLELE, CASECONTROL, FAMILY, HAPLOTYPE, HTSNP, INBREED, PSMOOTH. General routines in SAS such as MULTTEST and LOGISTIC are better known.

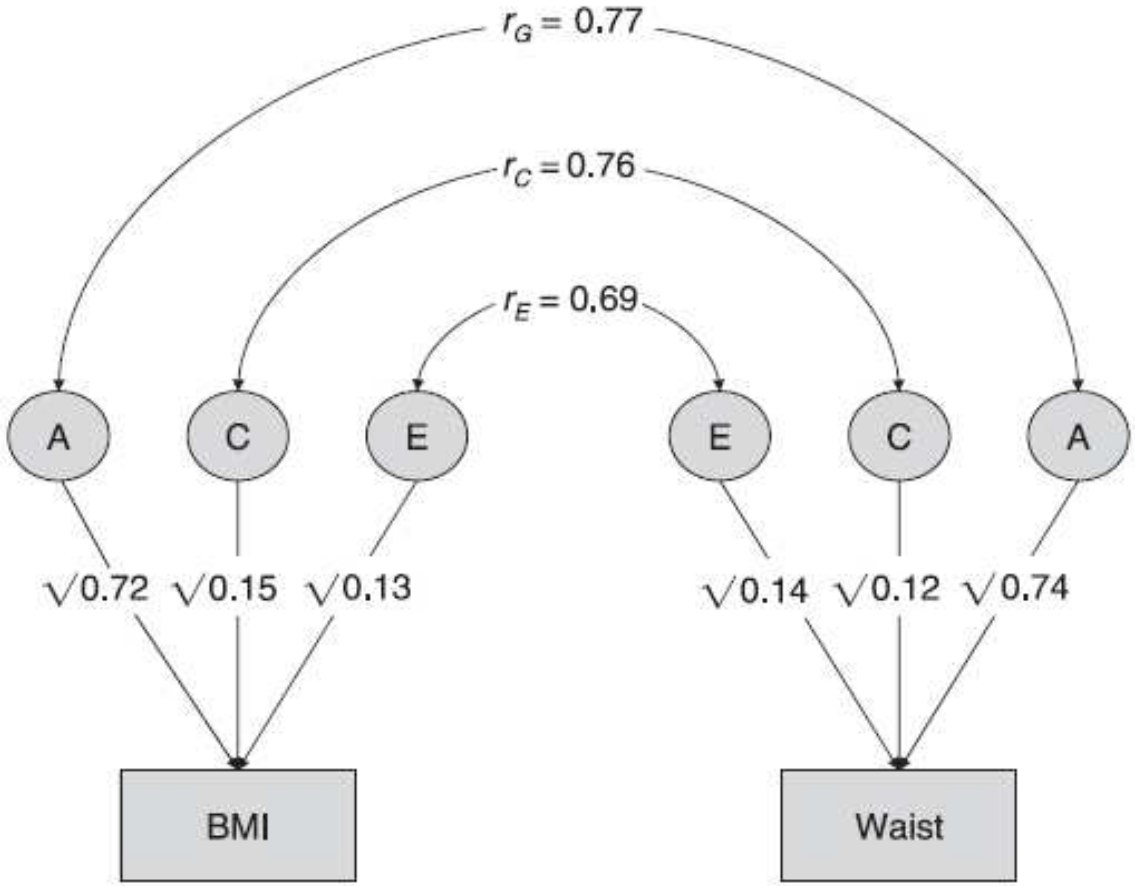
Stata

- Information available from http://www.stata.com
- The most popular software for epidemiologists, and researchers in other fields such as econometrics. Its facilities range from basic data management, computer graphics, programming, to methods for complex survey, longitudinal data analysis and multilevel models. Beside its simple but flexible syntax and comprehensive on-line documentation, it has many unique features such as frequency, probability and analytic weights.
- Support for databases, Internet, multiprocessor
- Stata Journal, Bulletin, a large repository and user community
- There are functions for descriptive analysis of markers, Hardy-Weinberg equilibrium, haplotype analysis and power of calculations.

R

- Linkage analysis: power (powerpkg, gap), check of relationships (Relcheck), pedigree analysis (identity, MasterBayes,multic), plot a genome scan (lodplot), survival models (kinship), meta-analysis (Leiden)
- Generic packages as given in ctv for genetics: genetics (Warnes. *R News*, **3**, 9-13, 2003), haplo.stats (Lake et al. *Hum. Hered.*, **55**, 56-65, 2003), gap (Zhao. J Stat Soft **23**(8):1-18, 2007), others (LDheatmap, luca, dgc.genetics), multiple testing: qvalue (multtest, twilight, locfdr,...), Bioinformatics: R News 6/5 (http://cran.r-project.org/doc/Rnews/Rnews_2006-5.pdf)
- Packages for genomewide association studies (GWAS), SNPassoc (Gonzalez et al. *Bioinformatics*, **23**, 644-645, 2007), GenABEL (Aulchenko et al. *Bioinformatics*, **23**, 1294-1296, 2007), SNPmatrix (Clayton and Leung, *Hum. Hered.*, **64**, 45-51, 2007), pbat2

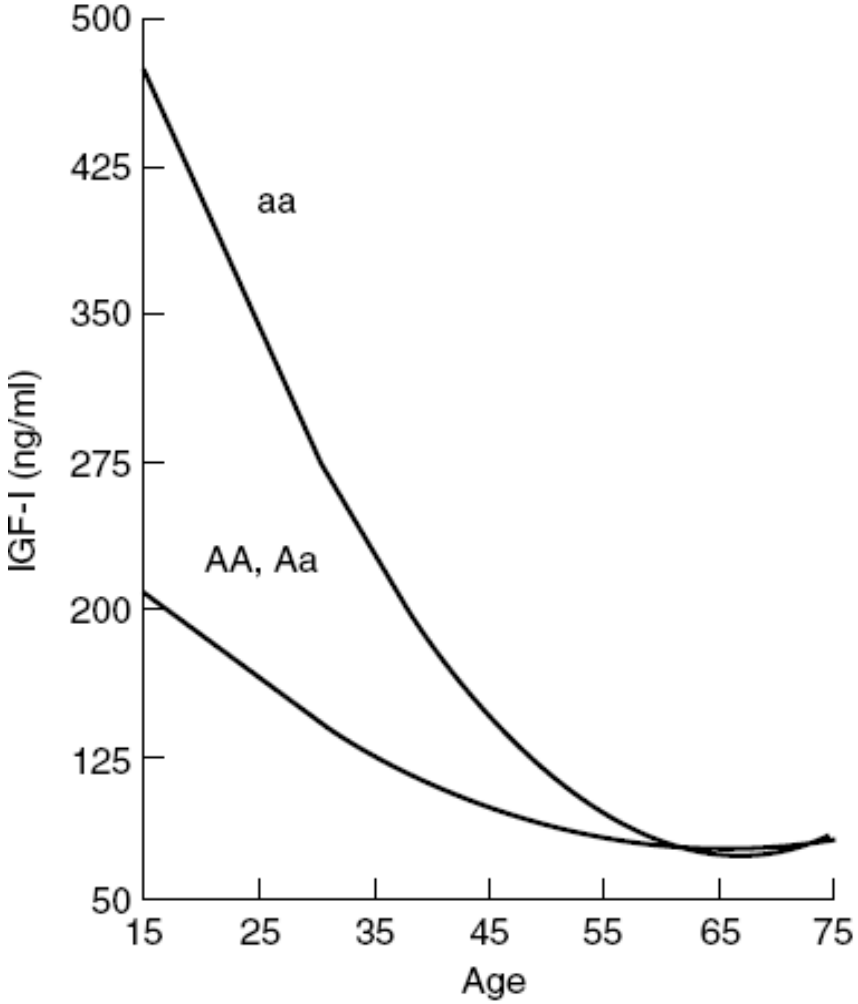
Estimation of heritability (Wardle et al. 2008)



Analysis of IGF1 in Mexican Americans

A study of 422 Mexican Americans from 24 pedigrees for insulin-like growth factor 1 (IGF1) showed evidence of a major gene influencing IGF1 levels. The estimated frequency of the A allele associated with lowered IGF1 levels was 0.54+/-0.05. Log-likelihood ratio tests revealed evidence for genotype by age interaction. Individuals with genotypes AA and Aa showed a less marked decline in IGF1 levels with age than that observed for the aa genotype.

(Blangero 2005)



Linkage of Huntington's disease

		Recombination fraction (θ)					
		0.0	0.05	0.1	0.2	0.3	0.4
Huntington's disease against G8	A	1.81	1.59	1.36	0.90	0.48	0.16
	V	6.72	5.96	5.16	3.46	1.71	0.33
	T	8.53	7.55	6.52	4.36	2.19	0.49
Huntington's disease against MNS		$-\infty$	−3.22	−1.70	−0.43	−0.01	0.07
Huntington's disease against GC		$-\infty$	−2.27	−1.20	−0.32	0.00	0.07
G8 against MNS		$-\infty$	−8.38	−3.97	−0.55	0.45	0.37
G8 against GC		$-\infty$	−2.73	−1.17	−0.08	0.14	0.08

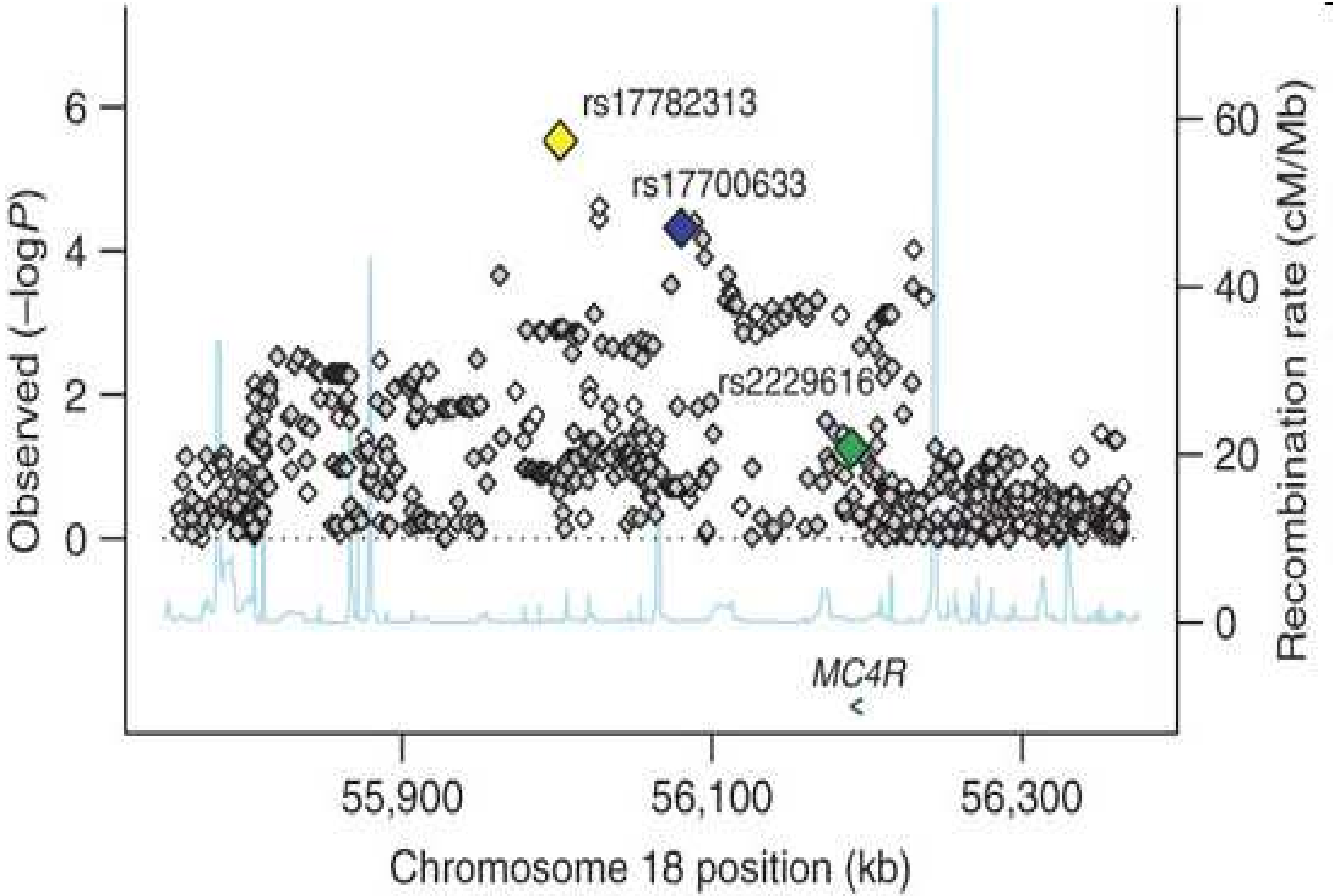
A, American pedigree; V, Venezuelan pedigree; T, total.

- The lod score between the Huntington's disease locus and G8 locus at chromosome 4 is 6.52, with 99%CI 0-10cM. However there is no evidence of linkage with MNS and GC loci. (Gusella et al. 1983)

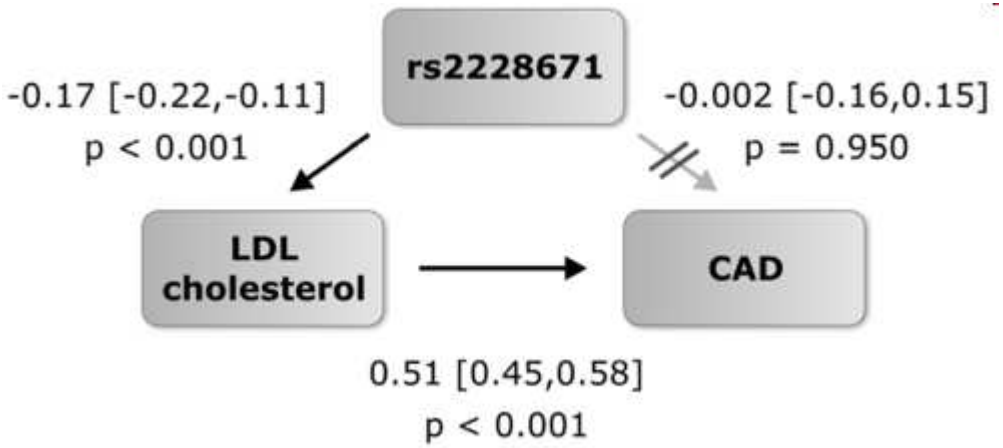
TCF7L2 and type-2 diabetes (Grant et al. 2006)

	Allele	Affected frequency	Control frequency	RR	Two sided *P*
All alleles of DG10S478					
Iceland (1,185/931)	0	0.636	0.724	0.67	2.1×10^{-9}
	4	0.005	0.002	2.36	0.12
	8	0.093	0.078	1.21	0.090
	12	0.242	0.178	1.48	4.6×10^{-7}
	16	0.022	0.015	1.53	0.076
	20	0.001	0.003	0.39	0.17
Denmark (228/539)	0	0.669	0.740	0.71	0.0048
	4	0.002	0.004	0.59	0.62
	8	0.070	0.048	1.49	0.091
	12	0.239	0.190	1.34	0.032
	16	0.020	0.018	1.12	0.78
USA (361/530)	-4	0.001	0.000	–	–
	0	0.615	0.747	0.54	3.3×10^{-9}
	4	0.003	0.004	0.73	0.72
	8	0.085	0.049	1.79	0.0029
	12	0.256	0.180	1.57	1.2×10^{-4}
	16	0.040	0.020	2.07	0.012
Allele X of DG10S478					
Iceland (1,185/931)	X	0.364	0.276	1.50 [1.31, 1.71]	2.1×10^{-9}
Denmark (228/539)	X	0.331	0.260	1.41 [1.11, 1.79]	0.0048
USA (361/530)	X	0.385	0.253	1.85 [1.51, 2.27]	3.3×10^{-9}
Combined	X	–	–	1.56 [1.41, 1.73]	4.7×10^{-18}

Association of *MC4R* and obesity (Loos et al. 2008)



Mendelian randomization



Causal relationship between LDL-C associated with rs2228671 and CAD (Mendelian Randomisation). In the structural equation model, carriage of the T allele at rs2228671 leads to lower LDL-C levels, and higher LDL-C levels lead to an increased risk of CAD. Given this, there is no additional direct path from rs2228671 to CAD risk, indicating that the functional pathway between the genetic variant at the *LDLR* gene locus and risk of CAD is through changes in LDL-C.

Linsel-Nitschke et al. (2008) PLoS ONE, 3(8):e2986

Challenges

- A full picture yet to piece together
- Data handling, e.g., intensity data
- Multiple testing
- Multivariate models
- Discovery and characterization in consideration of the pathways
- …

To wrap up

- The availability of DNA polymorphisms has for ever changed the scene of genetic study of human traits.
- It has offered scope and proved highly successful for interdisciplinary research of statisticians, computer scientists, molecular biologists and epidemiologists.
- There are still many problems to investigate traits specific to population under question.
- The field is highly dynamic; the course hopefully can guide our discussion over a range of issues many of which will be fruitful collaborations.

Readings and data descriptions

- General
  - Elston and Anne Spence. Stat Med 25:3049-3080, 2006
  - Balding. Nat Rev Genet 7:781-791, 2006
- Familial aggregation
  - Boyle and Elston. Biometrics 35:55-68, 1979
- Linkage analysis:
  - Gusella et al. Nature 306:234-238
  - Ohadi et al. Am J Hum Genet 64:165-171, 1999
- GWAS:
  - Pearson and Manolio. JAMA 299:1335-1344
  - Loos et al. Nat Genet 40:768-775, 2008
- Datasets to be used